

Использование фактов для поиска мнений в новостях

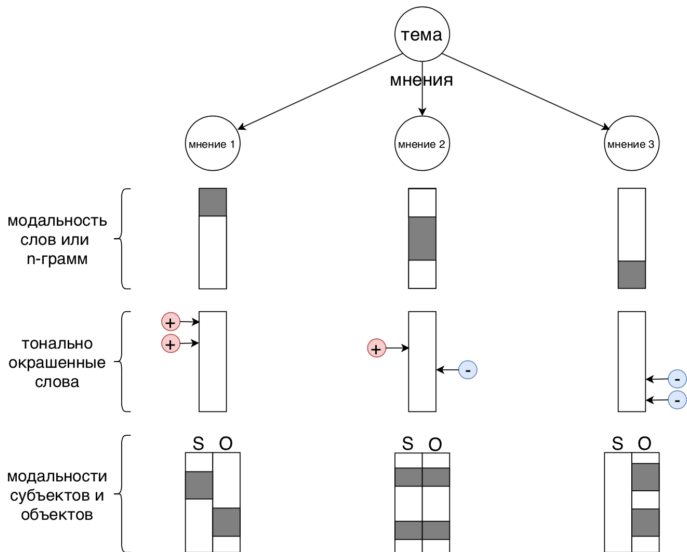
Фельдман Даниил

Научный руководитель: д.ф.-м.н., профессор В. А. Серебряков

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра прикладных проблем теоретической и математической физики

Москва, 2018

Определение мнения



Постановка задачи

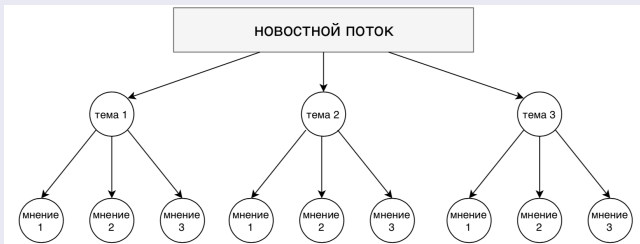
Дано

Корпус текстовых документов D на заранее выбранную тему и число мнений $|O|$

Задача

Распределить все документы в корпусе на $|O|$ групп, в которых все документы выражают некоторое мнение.

Применение



Представление текста

Последовательность троек $(w_i, o_i, d_i)_{i=1}^n \in W \times O \times D$

$$p(w|d) = \sum_{o \in O} p(w|o)p(o|d) = \sum_{o \in O} \varphi_{wo}\theta_{od}$$

предполагаем, что $p(w|o, d) = p(w|o)$

Введем матрицы $\Phi = \{\varphi_{wo}\}_{W \times O}$, $\Theta = \{\theta_{od}\}_{O \times D}$, $F = \{p(w|d)\}_{W \times D}$

Функция правдоподобия

$$L((w_i, d_i)_{i=1}^n; \Phi, \Theta) = \prod_{i=1}^n p(w_i, d_i) = \prod_{d \in D} \prod_{w \in W} p(w|d)^{n_{dw}} p(d)^{n_{dw}} \rightarrow \max_{\Phi, \Theta}$$

SyntaxNet — предобученная нейросеть поверх TensorFlow, поддерживает 40 языков, включая русский.

Вход:

- список предложений

Выход, для каждого слова в каждом предложении:

- id (порядковый номер слова в предложении)
- id родительского слова (0 для корня)
- исходное слово
- нормальная форма
- часть речи: NOUN, VERB, ADJ, ADV, ...
- отношения с другими словами: nsubj, dobj, conj, cc, nmod, ...

Типы триплетов

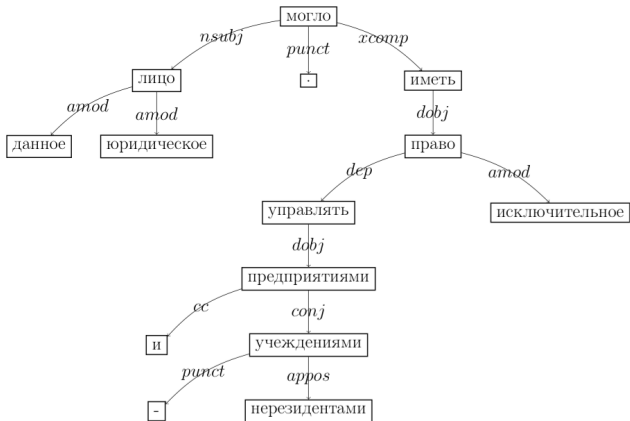
- Субъект-глагол-объект
- Субъект-причастие-объект
- Субъект-"есть"-объект
- Субъект-"есть"-прилагательное

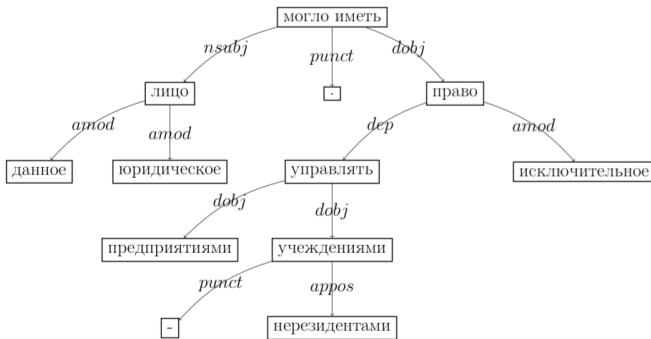
Поиск триплетов

1. Обработка однородных членов
2. Обработка имен и местоимений
3. Обработка сложных глаголов
4. Поиск основы триплета (глаголов, причастий, прилагательных и дополнительных существительных)
5. Дополнение основ до триплетов

Синтаксическое дерево

Данное юридическое лицо могло иметь исключительное право управлять предприятиями и учреждениями-нерезидентами





Модальности - субъекты и объекты

$$n_{dw} = \sum_{(s,p,o) \in T_d} [s = w], w \in W^m$$

Разреживающий регуляризатор

$$R(\Phi, \Theta) = -\beta_0 \sum_{o \in O} \sum_{w \in W} \beta_w \ln \varphi_{wo} - \alpha_0 \sum_{d \in D} \sum_{o \in O} \alpha_o \ln \theta_{od}$$

Максимизирует разницу между моделируемыми распределениями φ_o и θ_d и заданными распределениями $\beta = (\beta_w)_{w \in W}$, $\alpha = (\alpha_o)_{o \in O}$.

Сглаживающий регуляризатор

$$R(\Phi, \Theta) = \beta_0 \sum_{o \in O} \sum_{w \in W} \beta_w \ln \varphi_{wo} + \alpha_0 \sum_{d \in D} \sum_{o \in O} \alpha_o \ln \theta_{od}$$

Регуляризатор декоррелирования

$$R(\Phi, \Theta) = -\gamma \sum_{o \in O} \sum_{o' \in O \setminus o} \sum_{w \in W} \varphi_{wo} \varphi_{wo'}$$

Увеличиваем расстояние между φ_o

Гипотезы

- 1 Построение вероятностных моделей оправдано
- 2 Использование SPO триплетов дает прирост качества

Описание данных

Два размеченных корпуса:

- 1 100 текстов про национализацию предприятий в ЛНР и ДНР
- 2 220 текстов про решение Трампа выйти из Парижского соглашения

Метрика качества

Корпус разделен на кластеры $D = D_1 \cap D_2 \cap \dots \cap D_{|O|}$

$$PW(D) = \frac{\sum_{i=1}^{|O|} |\{c(d_1) = c(d_2) \mid (d_1, d_2) \in D_i \times D_i\}|}{\sum_{i=1}^{|O|} |\{(d_1, d_2) \in D_i \times D_i\}|}$$

TF-IDF

$$f(d) = \{\text{tf-idf}(w_i), w \in W\}$$

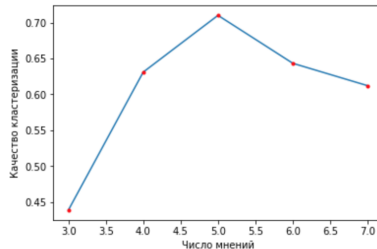
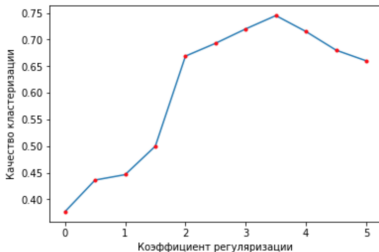
Среднее Word2Vec

$$d = (w_1, w_2, \dots, w_{n_d})$$

$$f(d) = \frac{1}{n_d} \begin{bmatrix} \text{Word2Vec}(w_1) \\ + \\ \text{Word2Vec}(w_2) \\ + \\ \dots \\ + \\ \text{Word2Vec}(w_{n_d}) \end{bmatrix}$$

Подбор параметров

	Корпус 1	Корпус 2
Число мнений	5	5
Число фоновых мнений	2	2
Минимальный tf	3	3
Вес субъектов	0.8	0.75
Коэффициент регуляризации	3.5	3.0



	$PW(D)$
TF-IDF	0.41
Word2Vec mean	0.42
ARTM	0.65
ARTM+SPO	0.74

Table: Предприятия ЛНР и ДНР

	$PW(D)$
TF-IDF	0.42
Word2Vec mean	0.39
ARTM	0.56
ARTM+SPO	0.62

Table: Парижское соглашение

Выводы

- 1 Лексических признаков недостаточно для кластеризации по мнениям
- 2 Использование SPO триплетов дает стабильный прирост качества в задаче поиска мнений

Спасибо за внимание!